

統計の利用【区間推定】

統計についてもう少し詳しい用語と使い方を学んでいきましょう。

(1) 平均値等を一定の幅をもたせて推定する～区間推定

①区間推定

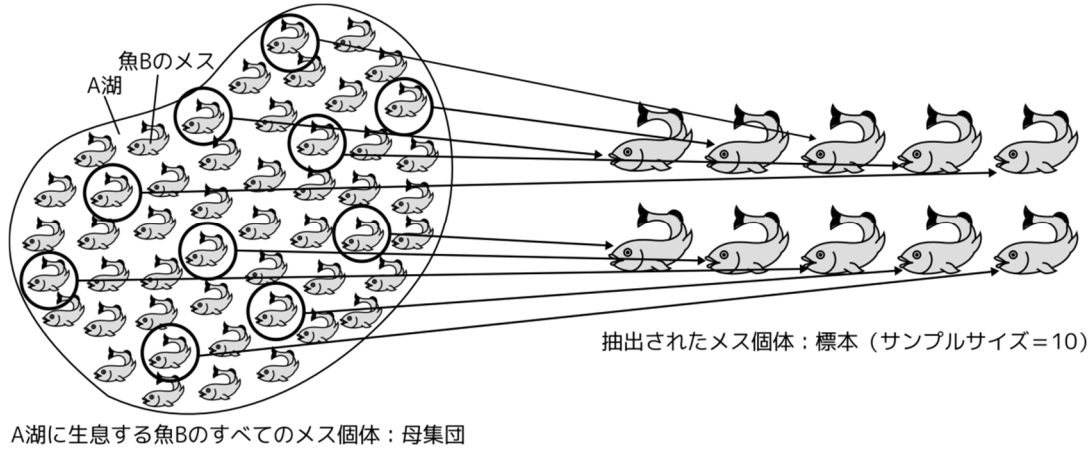
区間推定は、標本の統計量を元に、母集団の平均などを、幅（区間）をもたせて推定する統計学の手法です。この推定した幅を【3】といいます。例えば、母集団が大きい場合でも、区間推定を使えば、すべてのデータを測定しなくても、平均を推定することができるのです。今回の授業では、母平均（母集団の平均）の信頼区間の求め方について、具体例とともにその概要を理解しましょう。

例えば、今回あなたは次のようなことを知りたいものとし

【例1】ある湖（A湖）に生息している、ある魚（魚B）のメス1個体が、産卵期に抱卵している（お腹に持っている）卵の平均的な個数を知りたい

②母集団から標本を抽出する

例1のように、A湖に生息している魚Bのすべてのメス個体について抱卵数を知りたい場合、「A湖に生息している魚Bのすべてのメス個体」のことを【4】といいましたね。日本全国とは言いませんが、1つの湖といえども母集団は数が多すぎるので、全部を調べるわけにはいきません。そこで、図1のように、母集団の中から限られた数のデータを取ってきます。



▲図1 母集団から標本を抽出する

母集団の中から抽出したデータを【6】(サンプル), 標本のデータの数【7】あるいは**標本の大きさ**と呼びます。標本を抽出するときには、【8】に(ランダムに)抽出すること(=【9】)(ランダムサンプリング))が重要です。例えば、湖の1か所のみで個体を採取すると偏ったデータになることも予想されるため、湖の各所から数個体ずつ採取するなどの工夫をこらす必要があるのです。

こうして、A湖から魚Bのメス10個体を捕獲し、解剖して抱卵数を調べた結果は下の表2のようになったものとします。

▼表2 A湖に生息する魚Bのメス1個体あたりの抱卵数(サンプルサイズ=10)

No	1	2	3	4	5	6	7	8	9	10
抱卵数	47	51	49	50	49	46	51	48	52	49

③母平均を推測する～区間推定と信頼区間

さきほどの標本における魚Bのメス1個体あたりの抱卵数の平均(これを【1.

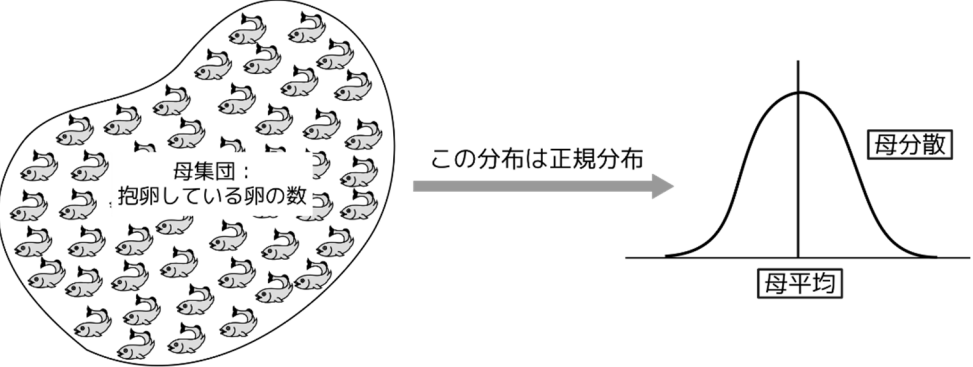
は、49.2個でした。ただ、この平均を、A湖に生息している魚Bのすべてのメス個体の抱卵数の平均(これを

【2】という)するのは、いくらなんでも無理があります。このような場合は、**区間推定**の出番です。つまり、「**標本から推定すると、母平均はこの値からこの値の間に入るのではないか**」という形で推定を行うのです。

①区間推定の考え方

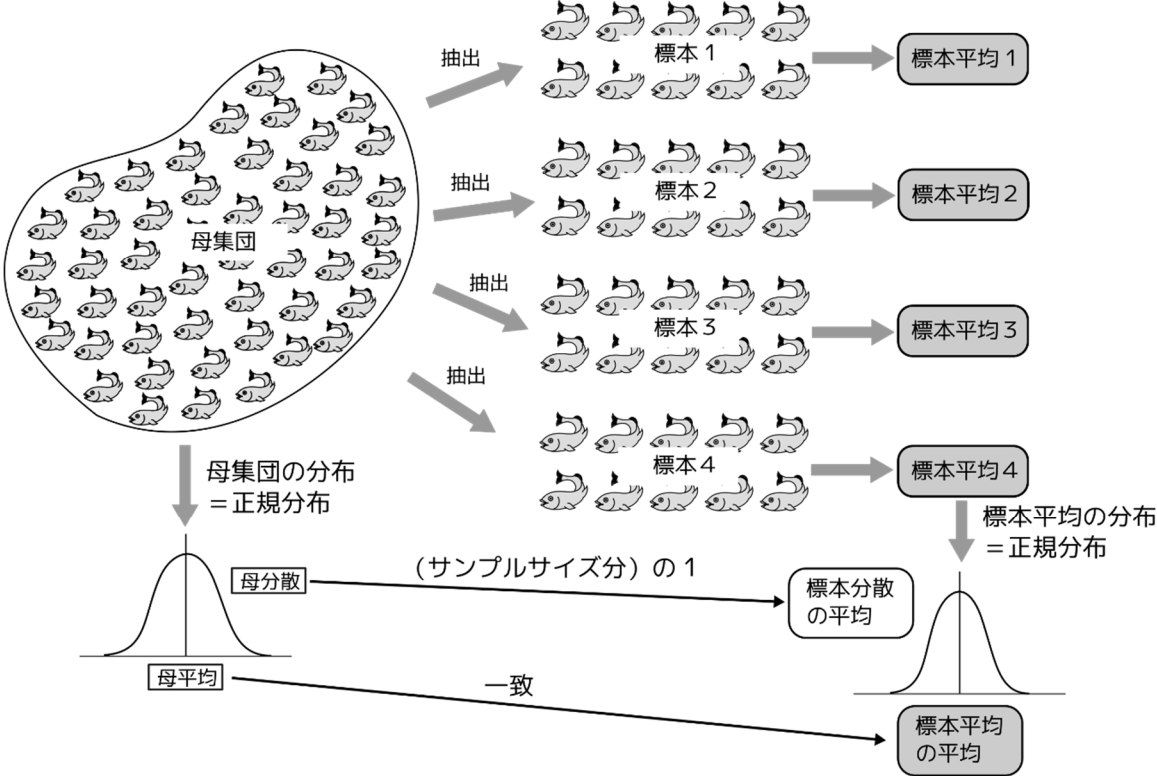
まず魚Bのメス1個体あたりの抱卵数を考えます。この分布が【3】にしたがっているものとし

▼図4 メス1個体あたりの抱卵数は正規分布にしたがっているものとする



次に図5のように、サンプルサイズを10として、母集団から10個のデータを抽出します。この平均を計算して、それを標本平均1とします。標本を戻して、もう一度10個のデータを抽出します。その平均を標本平均2とします。この作業を何回もくり返します。そうすると、図5のように、標本平均がたくさん求められます。これらの標本平均の分布を描いてみると、これも正規分布になります。このとき、標本平均の平均値は母平均に一致します。また、標本平均の分散は、母分散の「(サンプルサイズ)分の1」になります。図5では、サンプルサイズは10なので、標本平均の分散は母分散の10分の1になります。

▼図5 母集団の分布と標本平均の分布



先ほどの魚の例に戻って計算してみましょう。

このデータから計算すると、平均は（ 49.2 ）、分散は（ 3.51 ）、標準偏差は（ 1.87 ）となります（下表）。

▼表 データ集

1	47	-2.2	4.84
2	51	1.8	3.24
3	49	-0.2	0.04
4	50	0.8	0.64
5	49	-0.2	0.04
6	46	-3.2	10.24
7	51	1.8	3.24
8	48	-1.2	1.44
9	52	2.8	7.84
10	49	-0.2	0.04
合計	492	2乗の合計	
平均	49.2	分散	
		標準偏差	

*分散と不偏分散

たった今、不偏分散を求める式を紹介しましたが、これは分散を求める式の「サンプルサイズ」（前回の授業では「データ数」としていました）の部分で「サンプルサイズ-1」にしたものです。分散の式は次のようなものでしたね。

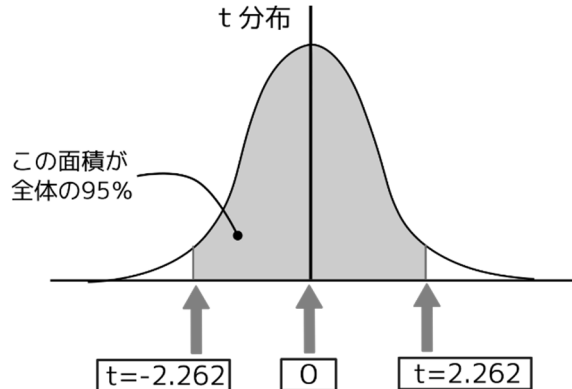
$$\text{分散} = \frac{(\text{データ} - \text{平均値})^2 \text{ の総和 }}{\text{（サンプルサイズ）}}$$

さて、これで、信頼区間を求めることができます。信頼区間は次の数式で求められます。

$$\text{信頼区間} = \text{標本平均} \pm t \times \text{標準誤差}$$

「t」は信頼区間で決められた分布の面積が95%となるような数値です。この「t」の具体的な値を決めるために、正規分布に似た、つりがね型の **t 分布** を使います。t 分布はサンプルサイズによって、少しずつ形が変わります。

▼図9 サンプルサイズが10のときのt分布



サンプルサイズが10のときのt分布は図9のようになります。ここで面積が95%になるようなtの値を決めます。図9が示すように、-tからtまでで挟まれた面積が全体の95%になります。なお、tの値は、あらかじめ用意されたt分布表（次ページの表10）を使って調べることができます。サンプルサイズが10のとき、-tからtまでで挟まれた面積が全体の95%になるようなtの値は2.262となります。これで、信頼区間が計算できます。

・95%信頼区間

$$\begin{aligned} \text{信頼区間} &= \text{標本平均 } 49.2 \pm t \text{ の値 } 2.262 \times \text{標準誤差 } \sqrt{3.51 \div 10} \\ &= 49.2 \pm 2.262 \times 0.593 \\ &= 47.86 \sim 50.54 \end{aligned}$$

このように、47.86 から 50.54 が95%信頼区間になります。この意味は「母平均が95%の確率（当たる確率95%）で47.86 から 50.54 までの間に含まれる」ということです。

【問い】同様に、信頼度99%の場合の信頼区間を求めてみよう。なお、サンプルサイズが10のとき-tからtまでで挟まれた面積が全体の99%になるようなtの値は3.250となります。

▼表10 t分布表

自由度	確率95%	確率99%	自由度	確率95%	確率99%
1	12.706	63.657	18	2.101	2.878
2	4.303	9.925	19	2.093	2.861
3	3.182	5.841	20	2.086	2.845
4	2.776	4.604	21	2.080	2.831
5	2.571	4.032	22	2.074	2.819
6	2.447	3.707	23	2.069	2.807
7	2.365	3.499	24	2.064	2.797
8	2.306	3.355	25	2.060	2.787
9	2.262	3.250	26	2.056	2.779
10	2.228	3.169	27	2.052	2.771
11	2.201	3.106	28	2.048	2.763
12	2.179	3.055	29	2.045	2.756
13	2.160	3.012	30	2.042	2.750
14	2.145	2.977	40	2.021	2.704
15	2.131	2.947	60	2.000	2.660
16	2.120	2.921	120	1.980	2.617
17	2.110	2.898	∞	1.960	2.576

＊自由度について

この場合、**自由度**とはサンプルサイズから1を引いたものです。例えば、サンプルサイズが10のとき、自由度は9になります。なお、この表に自由度が見当たらないときは、近い自由度の値を使って下さい（例えば、自由度32であれば近い値である自由度30を使えばOK。また、300や500のときには、無限大（∞）の値を使って下さい。

●まとめ

- ・ それについて知りたいと思う、全体のデータを**母集団**という。
- ・ 母集団からいくつか取りだしたデータを**標本**という。
- ・ 標本内のデータの個数を**サンプルサイズ**という。
- ・ よい標本を取り出すには、**無作為**（ランダム）に**抽出**することが必要。
- ・ 母平均は標本の平均で推定できる。
- ・ 母分散は標本の**不偏分散**で推定できる。
- ・ **不偏分散** = $((\text{データ} - \text{平均値})^2)$ の総和 ÷ (サンプルサイズ - 1)
- ・ 母平均が95%の確率で含まれているような範囲を **95%信頼区間**という。
- ・ **信頼区間** = 標本平均 ± t × 標準誤差
- ・ **標本平均の標準偏差（標準誤差）** = 不偏分散 ÷ サンプルサイズ
- ・ t の値は、確率（95%や99%）と**自由度**によって変わるので、**t 分布表**を見る。
- ・ サンプルサイズから1を引いたものを**自由度**という。

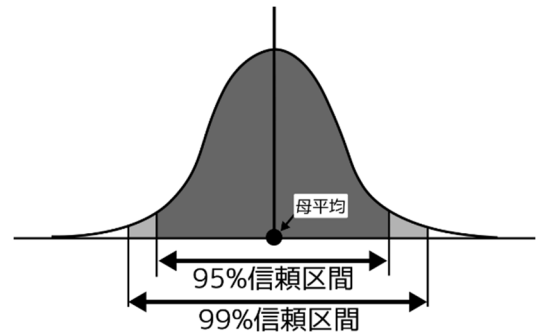
【補足】区間推定の基準となる「信頼度」

母集団から抽出した標本を使って区間推定をする際に、その区間の幅を決めるための基準が必要です。信頼区間の幅を決める基準となるのが「**信頼度**」（信頼水準・信頼係数）というものです。

信頼度として慣例的によく用いられる基準は、次の2種類です。

・ **信頼度：95%、99%**

この信頼度という基準を用いて、例えば右図のように正規分布であれば平均を中心とした区間の幅を決めるのでしたね。信頼度95%ならば、残りの5%にはどういう意味があるのでしょうか。1から信頼度を引いた値を「**有意水準**」といい、記号では α がよく用いられます。信頼度が95%ならば有意水準は5%となります。有意水準は危険率とも呼ばれるもので、いわば間違った答えを出してしまう割合です。例えば、有意水準5%というのは $5\% = 1/20$ なので、同じことを20回やったら1回ぐらいは間違った答えを出してしまうという程度を示しています。有意水準1%なら $1\% = 1/100$ なので、100回やれば1回ぐらいは間違うという程度になります。



【補足】信頼区間が意味するものとは

右図を見て下さい。縦の点線は母集団の真の平均を表し、横棒1本で1回の標本抽出での区間推定の結果を示しています。20本ある各横棒の中央の点が標本平均、実線が信頼度95%の信頼区間、点線が信頼度99%の信頼区間を表しています。信頼度が高いほど信頼区間は広い幅をとります。この図では同じ標本抽出を20回やったときに1回だけ95%信頼区間内に真の値が入っていない回（図中の○印）があることがわかります。このように、信頼区間とは、信頼度という基準で区間幅を推定したとき、同じ条件で標本抽出を何回か繰り返せば、区間内に真の値が含まれる回数はこの程度だということを意味する幅を示しています。

